

The effects of auditory-visual vowel identification training
on speech recognition under difficult listening conditions^a

Carolyn Richie^{b, c} and Diane Kewley-Port

Department of Speech and Hearing Sciences

Indiana University

Bloomington, Indiana 47405

Running head: Auditory-visual vowel identification training

- a) Portions of these data were presented at the November 2003 and May 2005 meetings of the Acoustical Society of America.
- b) Electronic mail: criche@butler.edu
- c) Current address: Communication Disorders, Butler University, Indianapolis, Indiana 46208

ABSTRACT

Purpose: The effective use of visual cues to speech provides benefit for adults with normal hearing in noisy environments and for adults with hearing loss in everyday communication. The purpose of this study was to examine the effects of a computer-based, auditory-visual vowel identification training program on sentence recognition under difficult listening conditions.

Method: Normal-hearing adults were trained and tested under auditory-visual conditions, in noise designed to simulate the effects of a hearing loss. After initial tests of vowel, word, and sentence recognition, one group of participants received training on identification of 10 American English vowels in CVC context. Another group of participants received no training. All participants were then retested on vowel, word, and sentence recognition.

Results: Improvements were seen for trained compared to untrained participants, in auditory-visual speech recognition under difficult listening conditions, for vowels in monosyllables, and keywords in sentences.

Conclusions: Results from this study suggest benefit may be gained from this computer-based auditory-visual vowel identification training method.

I. INTRODUCTION

Numerous technological devices have been designed to ameliorate the effects of hearing loss. However, devices such as hearing aids and cochlear implants do not fully restore normal hearing. Speech understanding can thus be a source of difficulty for many listeners with hearing loss, even with assistive devices. Face-to-face communication involving auditory-visual speech perception has been shown to improve speech understanding for many of these persons. The acoustic and optical signals carry different information regarding the speech act; they are sometimes complementary rather than redundant (Grant & Walden, 1995; MacLeod & Summerfield, 1987). The benefit gained by using visual cues to speech, in addition to acoustic cues, has been amply demonstrated for people with normal hearing (Erber, 1969; O'Neill, 1954; Sumbly & Pollack, 1954) and people with hearing loss (Grant, Walden, & Seitz, 1998; Prall, 1957). Nonetheless, there exists great individual variability in the ability to use visual cues associated with speech (Erber, 1969; Grant et al., 1998; Heider & Heider, 1940), and it remains unclear whether and how this ability can be significantly improved with training.

The high levels of individual variability and the difficulty associated with increasing the ability to use visual cues in speech recognition has led some to argue that the ability is innate and cannot be learned. Others maintain that the ability to use visual cues to speech recognition can be improved if appropriate training methods are used (see Gagné, 1994, for a discussion, as well as White, Dancer, & Burl, 1996). Improved auditory-visual speech perception abilities may be very useful because they may lead to better understanding in difficult listening conditions for adults with hearing loss.

While some conflicting evidence exists, published studies on adults with severe to profound hearing loss suggest that experience of reduced auditory input may indeed be related to superior speechreading ability. Superior speechreading skill for listeners with hearing loss has been documented both for individuals (*experts* such as AA documented by Andersson & Lidestam, 2005) and groups of individuals with hearing loss. Tillberg, Rönnberg, Svärd, and Ahlner (1996) compared speech recognition in hearing aid users differing in terms of the age of onset of hearing loss and the degree of hearing loss. They showed that adults with earlier-onset hearing loss who had experienced hearing loss for a greater portion of their lives outperformed adults with later onset hearing loss, on visual-only tests of word and sentence recognition. Bernstein, Demorest, and Tucker (2000) examined visual-only speech recognition by normal-hearing listeners and listeners with severe to profound hearing loss, for CV monosyllables, words and sentences. Bernstein et al. showed that performance was enhanced for adults with hearing loss when compared to adults with normal hearing, in terms of sensitivity to phonetic information in the various materials. They claimed, "... the necessity to perceive speech without hearing can be associated with enhanced visual phonetic perception in some individuals" (p. 233). Recently, Mohammed et al. (2005) designed a speechreading test expressly for deaf adults; it included vocabulary and syntax familiar to deaf adults and allowed for responses to be given via picture choice in order to eliminate familiarity with written English as a confounding variable. The speechreading test assessed word, sentence, and connected speech understanding, and was administered in the preferred communication mode for each participant. When this new test of speechreading skill was administered, Mohammed et al. found that deaf adults were better speechreaders than normal hearing

adults. Taken together, this research on adults with severe to profound hearing loss suggests that exposure to the speech signal under difficult listening conditions may be associated with enhanced speechreading abilities.

Several researchers have attempted to improve auditory-visual speech perception in adults with normal hearing and adults with hearing loss through training. These studies have taken a variety of approaches to speech perception training, and evaluated speech recognition for vowels, consonants, syllables, words and phrases (Bernstein, Demorest, Coulter, & O'Connell, 1991; Binnie, 1977; Black, O'Reilly, & Peck, 1963; De Filippo, 1984; Eberhardt, Bernstein, Demorest, & Goldstein, 1990; Gesi, Massaro, & Cohen, 1992; Lesner, Sandridge, & Kricos, 1987; Massaro, Cohen, & Gesi, 1993; Montgomery, Walden, Schwartz, & Prosek, 1984; Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). No study to our knowledge, however, has examined the effects of auditory-visual vowel identification training on speech recognition, and this gap provided motivation for the present investigation. Specifically, this study examined how auditory-visual vowel identification training affected syllable, word, and sentence identification. In contrast to previous research that involved training on consonant recognition (Lesner et al., 1987; Massaro et al., 1993; Walden et al., 1981; Walden et al., 1977), this study focused on vowel identification training given that vowel information is important to sentence understanding (Burkle, Kewley-Port, Humes, & Lee, 2004; Cole, Yan, Mak, Fanty, & Bailey, 1996; Kewley-Port, Burkle, & Lee, 2007).

A. Auditory-visual speech perception training using vowels

There is limited research related to auditory-visual speech perception training and vowels. Though some aural rehabilitation training programs advocate visual vowel identification training (e.g., Tye-Murray, 2004), they are few in number and sometimes unclear in their descriptions of specific training methods to be used (see Heider & Heider, 1940; Hutton, 1959). More often, in lipreading training books there are simply notes on which vowels are *easy* versus *difficult* to recognize. Support for a vowel-based approach to auditory-visual speech perception training can be gathered from related research on the informational value of vowels to speech recognition.

While consonants seem to be much more important than vowels to overall word recognition when reading printed English (Lee, Rayner, & Pollatsek, 2001), research in spoken English suggests that vowels are at least as important as consonants in speech perception. Diehl, Kluender, Foss, Parker, and Gernsbacher (1987), for example, showed that vowels are an important source of information in auditory recognition of consonants. In their study of CVC syllables containing 10 different American English vowels, the participants were asked to monitor the words for initial consonant. The results showed that accurate consonant identification is largely dependent upon recognition of the following vowel. Diehl et al. concluded that vowels are more easily identified than consonants and as such act as “islands of reliability in the speech stream” (p. 569).

Cole et al. (1996) also suggested that vowels contribute more information to overall auditory speech intelligibility than consonants. Young adults with normal hearing listened to sentences from the Texas Instruments / Massachusetts Institute for Technology (TIMIT) database (Garofolo et al., 1990). The participants identified words in unaltered sentences, sentences in which all of the vowels were replaced with noise

(consonants-only sentences), and sentences in which all of the consonants were replaced with noise (vowels-only sentences). The results showed that approximately twice as many words were recognized in the vowels-only compared to the consonants-only sentences. The location of the segment boundaries did not strongly affect the large difference in sentence intelligibility found between the vowels-only and consonants-only conditions.

Kewley-Port and colleagues (Burkle et al., 2004; Kewley-Port et al., 2007) replicated and extended Cole et al.'s study using both young adults with normal hearing and elderly adults with hearing loss. The results showed that word intelligibility for vowels-only sentences was significantly greater than for consonants-only sentences, by 14% to 40% across groups. Kewley-Port et al. concluded that vowels contribute more information than consonants to auditory sentence intelligibility, for both young adults with normal hearing and elderly adults with hearing loss.

Support for a vowel-based approach to auditory-visual speech perception training also can be drawn from one of the earliest empirical studies of lipreading, done by Heider and Heider (1940), who addressed the issue of the relative contribution of vowels and consonants to speech understanding. Heider and Heider suggested that recognition of vowels may be more important than consonants for learning how to lipread. They pointed out that the visual differences among some consonants are often non-existent (an idea similar to visemes, a term introduced later by Fisher (1968)). Heider and Heider noted,

...that a certain consonant belongs, for instance, to the group of *m*, *p*, and *b*, one can see without much training – and no amount of training makes it possible to distinguish accurately between these three consonants. On the

other hand with vowels, where the differences are much more gradual, one can learn to make finer and finer differentiations (p. 141).

They tested children with hearing loss who attended the Clarke School for the Deaf for more than one year and participated in lipreading training. A significant correlation was obtained between vowel and sentence lipreading abilities, while a similar consonant test showed less reliability and lower correlation with general lipreading ability. Heider and Heider argued that what distinguishes good lipreaders from poor lipreaders is lipreading ability for vowels, not consonants.

Another argument in support of vowel-based auditory-visual speech perception training is that vowels often remain relatively audible compared to consonants, for the person with hearing loss. Vowels usually have higher intensity and longer duration, relative to consonants. As Lesner et al. (1987) noted, people with high-frequency sensorineural hearing loss typically demonstrate consonant confusions but less difficulty with vowel identification. This being the case, auditory-visual vowel identification training may capitalize on an ability of the person with hearing loss, rather than something that causes difficulty (although the value in consonant information is duly noted). Hack and Erber (1982) wrote, “The vowel nucleus of a word often may be the only part clearly audible to the hearing-impaired listener. If the vowel is incorrectly identified the word and consequently the entire sentence may be misunderstood” (p. 100).

B. Auditory-visual cues to vowel perception

The auditory-visual cues for vowels are qualitatively different from the cues for consonants. By nature of their production vowels are more continuous than most consonants. As Ades (1977) noted, if a synthetic acoustic stimulus set is designed to

range from one phonemic category to another, there will be more just noticeable differences within the vowel set than in the consonant set. The visual cues to vowels, often described via articulatory phonetic features, are graded rather than categorical (McGrath, Summerfield & Brooke, 1984). For monophthongal (steady-state) English vowels, for example, Jackson, Montgomery, and Binnie (1976) found that an extended-rounded feature, a vertical lip separation feature, and a feature related to overall area of maximum lip opening are related to visual vowel identification

While visual consonant perception often is discussed in terms of visemes, some researchers suggested that vowels are not likewise indistinguishable. Sanders (1993), for example, noted that theoretically there are no homophenous vowels, and Berger (1970) observed that most lipreading instruction books treat each vowel as if it were an independent viseme. Although theoretically there may be no homophenous vowels (according to Sanders), there are nevertheless some vowels that look very similar to each other, or are often confused. As Hack and Erber (1982) noted, however, instead of forming distinct clusters with equally confusable members as consonants do, vowels are typically confused with adjacent vowels in the articulatory space.

Because vowel viseme categories may be organized differently than consonant viseme categories, vowels may present potential for learning to better recognize speech under difficult listening conditions. It is possible that the visual cues to vowels may aid overall auditory-visual speech perception similarly to the way that the acoustic cues to vowels are advantageous in noise replacement studies (Burkle et al., 2004; Cole et al., 1996; Kewley-Port et al., 2007). Because of the fact that vowels are produced in a manner that is continuous rather than categorical, continuous articulatory movements

toward and away from vowel nuclei may provide important information for visually identifying surrounding consonants, just as the formant transitions toward and away from vowel nuclei have been shown to contribute useful information in the auditory recognition of speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

Although research exists on the vowel lipreading abilities of untrained participants, and a few aural rehabilitation programs advocate vowel-based lipreading training for listeners with hearing loss, studies on vowel-based auditory-visual speech perception training are not readily found. The purpose of this study was to determine whether vowel identification training can improve auditory-visual speech recognition for words and sentences, under difficult listening conditions

II. METHODS

A. Participants

Fourteen adults between 19 and 28 years of age were tested in this experiment (mean age = 22.8 years; $SD = 3.3$ years). Seven participants (3 females, 4 males) were randomly assigned to the *untrained* group and 7 participants (4 females, 3 males) were assigned to a *vowel trained* group. All participants were native speakers of general American English, and were paid for their participation. Their hearing was found to be normal, evidenced by audiometric pure-tone thresholds ≤ 20 dB HL at octave frequencies from 250 to 4000 Hz. Vision also was tested using *The MNRead Acuity Charts* (Mansfield, Legge, Luebker, & Cunningham, 1994). This test measures reading acuity (the smallest print that can just be read) from a distance of approximately 30 cm, and maximum reading speed (the reading speed when performance is not limited by print

size). All participants in the study had normal or corrected-to-normal binocular vision, as defined by acuity of 20/25 or better and reading speed within normal limits.

B. Stimuli

Stimuli in this experiment consisted of short movie clips from Sensimetrics' (2001) CD-Rom, and Bernstein and Eberhardt's (1986) videodisk. Stimuli from Bernstein and Eberhardt were digitized from their original format on laser disc. All stimuli were then converted to .avi format (digital graphics files) using a sample rate of 22,050 Hz for audio, and normalized to be the same area regardless of source. All visual stimuli were approximately 200 cm² in area when displayed on the computer monitor. Stimuli from Sensimetrics made use of a video resolution of 360 (w) X 480 (h) pixels, and stimuli from Bernstein and Eberhardt made use of 480 (w) X 360 (h) pixels resolution. Different dimensions were used to maintain the same overall area across stimuli but realistic, undistorted proportions. There were three types of movie clips used in the experiment: (1) monosyllables for the vowel identification pretest, training, and posttest; (2) words for the word recognition pretest and posttest; and (3) sentences for the sentence recognition pretest and posttest.

The stimuli used for the vowel identification tasks included ten English vowels /i, I, e, ε, æ, α, Λ, o, υ, u/ in isolated monosyllables. A total of 120 different movie clips were selected from both sources, and included six different talkers. The 10 vowels were presented in a variety of CVC contexts, some of which were real words such as *bib* /bɪb/, and some of which were phonotactically plausible non-words such as *beeb* /bib/. Between four and eight occurrences of each vowel were included in each set of 60 tokens. These

two blocks of 60 tokens each were used as both testing and training materials for the participants in the *vowel trained* group.

The word stimuli were a total of 40 different movie clips selected from both sources, and included six talkers. The words were composed of CVC, CVCC, and CCVC strings. The words were balanced for the 10 vowels under study; in the set of 40 words, there were four instances of each vowel. One block of 20 words was used in the pretest (two per vowel) and another block of 20 different words was used in the posttest (two per vowel). In this set of materials three talkers were seen in the pretest and six talkers were seen in the posttest, three of whom were familiar and three of whom were unfamiliar (i.e., not seen in the other training or test materials). It was expected that familiar talkers would be easier to understand than unfamiliar talkers in the posttest, for the *vowel trained* participants.

Finally, 40 Central Institute for the Deaf Everyday Sentences (Davis & Silverman, 1970) were selected from the Johns Hopkins Lipreading Corpus (Bernstein & Eberhardt, 1986), spoken by the male and the female talker. In the pretest, 10 sentences spoken by the female talker and a different 10 sentences spoken by the male talker were used. A different set of 20 such sentences was used in the posttest.

C. Protocol

A summary of the basic testing protocol is outlined in Table 1. All participants came for the first day of testing, which included hearing and vision screening as well as the series of pretests. Pretests were designed to assess the participants' untrained auditory-visual speech perception abilities. Three pretests were administered without feedback: a closed-set vowel identification test (vowels in CVC context); an open-set

monosyllable word recognition test; and an open-set sentence recognition test (Central Institute for the Deaf Everyday Sentences). The participants in the *vowel trained* group then returned for six training sessions spread over six or more days on the vowel identification task, with feedback. All participants returned for the final day of testing, which included posttests administered without feedback. The posttests mirrored the pretests but included different stimulus materials, including a vowel identification test, a word recognition test, and a sentence recognition test. The posttests were designed to assess the participants' auditory-visual speech perception abilities following six sessions of vowel training or an equivalent period of time with no training. All sessions lasted approximately one hour.

----- Insert Table 1 here -----

D. Pretest and posttest procedures

For all procedures, the participants were tested individually, seated at a computer in a sound-treated booth. The participants were instructed to sit a 'comfortable' distance from a 17 inch flat panel computer monitor (a distance of approximately 60 cm, not to exceed approximately 1 m). Each movie clip was presented only once. The participants were instructed to watch each movie and determine the appropriate response with no feedback. They were encouraged to guess if unsure.

In the vowel identification pretest and posttest, the participants were presented with 60 randomized short movie clips presented one at a time. After each movie was presented they were asked to match the vowel in the syllable to one of those in a list of keywords, and use a computer mouse to enter a response. Keywords were reviewed with

the participants, before the pretest and posttest, and produced by the experimenter. The experimenter then left the booth and the participants completed the test independently.

In the word recognition pretest and posttest, the participants watched 20 randomized short movie clips. The participants were instructed to say aloud what the speaker said, and then press the space bar on the keyboard to see the next movie. The experimenter was present in the sound booth for this test, and recorded each response. Responses were transcribed, and the participants' voices also were recorded using a small digital recording device (Olympus, model DW90) placed approximately 50 cm in front of the participants. The experimenter listened in quiet and phonetically transcribed the participants' responses. Responses were scored in terms of words correctly identified, and expressed in percent correct.

In the sentence recognition pretest, similar procedures were followed as for the word recognition test, and the participants watched movie clips of 20 sentences. The participants again responded aloud. The experimenter circled correct responses on an answer sheet and digitally recorded the participants' spoken responses. The 20 unique sentences were not randomized, with the female talker presented before the male talker. Twenty new movie clips were presented in the sentence recognition posttest. These two sets of 20 sentences were roughly counterbalanced within groups, such that half of each group in the experiment saw one set of 20 sentences in the pretest and the other 20 sentences in the posttest.

E. Training Procedures

The decision to train the participants for six hours total was based on results from a previous, similar training program designed by Walden et al. (1977). In their study of

training effects on visual recognition of consonants, Walden et al. discovered that most learning took place during the first five to six hours of training (out of 22 hours). In this study the six one-hour training sessions consisted of the vowel identification task, as described above, with feedback after each response. Two training blocks of 60 movie clips each were alternated (the same two blocks as used for the vowel identification pretest and posttest), until the participants completed six blocks total. Instructions similar to those administered in the pretest were repeated. Feedback text on the computer screen indicated whether the response was correct, or if incorrect, the correct response was given.

F. Listening conditions

Because participants with normal hearing were tested in this experiment, background noise was added to the audio channel. Shaped noise was used to simulate a hearing loss, so that future work could make comparisons between results from the present study for adults with normal hearing and adults with hearing loss. Noise-masked thresholds of the normal-hearing participants in this study were matched within ± 5 dB to those of a young adult with hearing loss tested in a previous study; YHI#5 reported in Richie, Kewley-Port, and Coughlin (2003). Air-conduction pure tone audiometric thresholds for the right ear of that participant with hearing loss were tested in quiet. Test results for the octave frequencies 250 through 4000 Hz were as follows; 30, 40, 50, 55 and 55 dB HL, respectively. According to Humes, Dirks, Bell, and Kincaid (1987), noise-masking may simulate some of the effects of hearing loss for listeners with normal hearing, such as increased thresholds and decreased dynamic range.

G. Calibration

To obtain the shaped noise, uniform noise was generated by a waveform generator (Tucker Davis Technologies, model WG2) and shaped with cascaded IIR and FIR filters (Tucker-Davis Technologies, PF1 modules) to obtain the desired spectrum. The noise was low-pass filtered at 4300 Hz, with roll-off of approximately 44 dB / octave. This noise was found to have an overall level of 81 dB SPL on the linear scale as measured using headphones (model TDH-39) in a 6 cm³ coupler with a sound level meter (Larson-Davis, model 800B).

A calibration vowel was selected as follows. A sample of movie clips of three words spoken by each of the 12 talkers in the experiment was selected. Phonetically *low* vowels are known to have relatively higher intensity than phonetically *high* vowels. Three vowels varying in phonetic height (high, mid, low) were thus chosen for each speaker, to represent a range of intensities produced by that speaker. The rms energy in a 50 msec window around the peak amplitude of each vowel was calculated, and converted to decibels. The overall median peak energy of the 36 vowels was determined to be 71 dB (the range was from 66 to 76 dB). One of the vowels in one movie clip, the /æ/ in /væn/ spoken by a female talker from the Sensimetrics materials, also was found to be 71 dB. The vocalic portion of this syllable was then extracted and iterated for six seconds, and used as the calibration vowel. Using the default settings of the sound card (CardDeluxe) on the test computer, the calibration vowel was measured to be 91 dB SPL on the linear scale, using headphones (model TDH-39) seated in a 6 cm³ coupler.

H. Pilot experiment to determine appropriate signal-to-noise ratio

A separate pilot study was conducted to find a signal-to-noise ratio (SNR) that yielded 30 - 50% correct auditory-visual identification of the 10 vowels in monosyllables.

Scores in this range indicate that participants were able to perform the task above chance (10% correct), and that there was sufficient room for improvement. An initial pilot SNR of -30 dB was chosen based on results from Sumbly and Pollack (1954), who showed that at that SNR, participants with normal hearing correctly identified approximately 41% of words in closed sets of 128 - 256 words. Robert-Ribes, Schwartz, Lallouache, and Escudier (1998) also tested 21 participants with normal hearing on auditory-visual identification of seven French vowels at -18 dB SNR, and those participants achieved approximately 50% correct. The participants in this pilot study were tested under auditory-visual conditions at an SNR of -30 dB. The shaped noise yielded an overall sound pressure level of 81 dB SPL (see calibration above). Given that the calibration vowel was measured as 91 dB SPL, the signal level was attenuated 40 dB (Tucker-Davis Technologies, PA4 module) to 51 dB SPL, resulting in an overall SNR of -30 dB.

Seven adults with normal hearing not participating in the experiment were tested as pilot participants. These participants identified vowels in monosyllable context, for two blocks of 100 movie clips presented randomly, and with the same instructions for the vowel identification experiment above. Results from this pilot study yielded an average of 44% ($SD = 5.5$) correct identification of vowels in monosyllables under auditory-visual conditions with a SNR of -30 dB. Thus, a SNR of -30 dB was used and monitored throughout the experiment.

III. RESULTS

A. Training results

All 7 participants in the *vowel trained* group showed improvement, with minor fluctuations, over the course of training (see Figure 1). Improvement ranged from gains of 18 to 35 percentage points, averaged over the 10 vowels. The average pretest score over all 10 vowels was 38% correct, while the average posttest score over all 10 vowels was 67% correct. Thus, the group average improvement in vowel identification was 29 percentage points, from pretest to posttest.

----- Insert Figure 1 here -----

B. Vowel identification performance

Vowel identification abilities for the *untrained* and *vowel trained* participants in this experiment were compared, from pretest to posttest, using a repeated-measures analysis of variance (ANOVA). Scores for each participant expressed in terms of a percent correct average over all 10 vowels are shown in Table 2. Confusion matrices for the 10 vowels, averaged over all participants, can be seen in Figure 2. As expected, the ANOVA showed that there were significant main effects of group, $F(1, 12) = 5.73, p < .05$, and test, $F(1, 12) = 43.23, p < .001$. There also was a significant group X test interaction, $F(1, 12) = 26.56, p < .001$. Taken together, these results indicate that auditory-visual speech recognition for vowels in monosyllables improved significantly for the participants in the *vowel trained* group compared to the participants in the *untrained* group. While the participants in the *vowel trained* group were slightly poorer by 3.2 percentage points compared to the *untrained* participants in terms of vowel identification in the pretest, they superseded the *untrained* participants in the posttest, by far.

----- Insert Table 2 and Figure 2 here -----

C. Word Recognition performance

The word recognition performance of the participants in both groups was scored in terms of percent correctly identified words and compared from pretest to posttest. Word recognition results can be seen in Table 3. Because both groups performed similarly in terms of word recognition on the pretest, and because neither group showed remarkable change from pretest to posttest, these results were not subjected to further statistical analysis.

As part of the experimental design, however, the pretest stimuli included three talkers, whereas in the posttest stimuli there were six talkers; three familiar talkers and three unfamiliar talkers that the participants had not seen previously. The purpose of including unfamiliar talkers in the posttest was to examine whether improvements in auditory-visual speech recognition were related to talker familiarity for the *vowel trained* group. Unexpectedly, posttest scores for the participants in both groups were the same for familiar and unfamiliar talkers (see Table 3), suggesting that word recognition performance was not influenced by talker familiarity per se. This was unexpected given that the *vowel trained* group had more exposure to the familiar talkers from the pretest through training (albeit with different speech tokens). This result also suggests that, as a group, the unfamiliar talkers were no more easy or difficult to understand than the familiar talkers.

----- Insert Table 3 here -----

D. Sentence recognition performance

The sentence level auditory-visual speech recognition abilities of the participants in this study were measured using Central Institute for the Deaf Everyday Sentences (see Table 4). The participants were tested on different sentences in the pretest and posttest.

For each set, 50 keywords were scored in terms of percent correct and transformed to rationalized arcsine units (RAUs; Studebaker, 1985) for purposes of statistical comparison. The RAU transform reduces the compressive effect of the percent correct scale for extreme values (at either end of the scale).

----- Insert Table 4 here -----

The sentence identification scores in RAUs for both groups were compared, from pretest to posttest, using a repeated-measures ANOVA. Results of this test showed neither a significant main effect of group, $F(1, 12) = 4.07, p = .066$, nor test, $F(1, 12) = 2.15, p = .169$. However, a significant group X test interaction was obtained, $F(1, 12) = 8.10, p = .015$. Post hoc analysis using the Tukey HSD test produced only one significant comparison showing that performance in the *vowel trained* group was significantly better at posttest than pretest. Thus results showed that performance on sentence identification increased significantly by 10 percentage points, from pretest to posttest, for the participants in the *vowel trained* group, whereas the *untrained* group performance decreased from pretest to posttest by three percentage points. It should be noted that the participants in the *vowel trained* group demonstrated a range of training effects. Two participants showed negative or negligible changes from pretest to posttest, however, the other 5 participants in the *vowel trained* group showed improvements from 10 to 24 percentage points.

IV. DISCUSSION

The results from this study indicated that vowel-based auditory-visual speech perception training led to improvement in auditory-visual speech perception performance

for vowels in syllables, and sentences. Unexpectedly, no training effects were seen for word recognition in the posttest. Perhaps the larger number of different talkers in the word recognition posttest (n=6) compared to the pretest (n=3) may have led to the unexpected lack of training effects. While the word lists used in the word recognition pretest and posttest were balanced in terms of the vowels, perhaps a confounding factor such as lexical frequency or neighbourhood density was present.

Modest improvements were observed at the sentence level for the trained participants in this study. Previous work has examined the effects of consonant-based auditory-visual speech perception training on sentence-level speech understanding, by adults with both normal hearing and hearing loss. The following consonant-training studies were considered most comparable to the present vowel-training study in terms of training paradigm and scope of test materials.

Walden et al. (1981) studied the effects of consonant recognition training on auditory-visual speech recognition by adults with hearing loss. Thirty-five adult males with hearing loss were enrolled in an aural rehabilitation program; 15 received the standard two-week aural rehabilitation program, 10 received an additional seven hours of auditory consonant training, and 10 received an additional seven hours of visual consonant training. The results showed that all three groups improved significantly on auditory-visual sentence recognition from pretest to posttest; the auditory training group by 28 percentage points, the visual training group by 23 percentage points, and the standard aural rehabilitation group by 10 percentage points.

It should be noted, however, that the same materials were used in the sentence recognition pretest and posttest in that study. Though Walden et al. (1981) argued that

this should not have affected the test results, other studies have demonstrated that such improvements may be due to increased familiarity with test materials and test situation, rather than learned generalizable skill in auditory-visual speech perception (Eberhardt et al., 1990; Squires & Dancer, 1986; Van Tasell & Hawkins, 1981; Warren, Dancer, Monfils, & Pittenger, 1989). Thus some caution must be exercised in interpreting the Walden et al. (1981) results.

Lesner et al. (1987) trained 30 young adults with normal hearing on a visual consonant recognition task; 10 received 14 hours of training (with auditory-visual feedback), 10 received seven hours of pseudo-training (exposure without feedback), and 10 did not receive training. Lesner et al. used training materials similar to those reported in Walden et al. (1977), presented visual-only. Visual-only posttest performance on a set of Central Institute for the Deaf Everyday Sentences different from those used in the pretest showed no significant differences between groups – in fact, all groups had lower speech recognition scores on the posttest than the pretest. The authors concluded that the lipreading training method employed in their study was not effective in improving visual only speech recognition performance for sentences.

Massaro et al. (1993) tested six females with normal hearing in six conditions combining speed of presentation (normal, fast) and modality (auditory, visual, auditory-visual). Testing was done at the beginning of the experiment, after each of five training tasks, and several weeks later, for CV syllables, monosyllable words, and Central Institute for the Deaf Everyday Sentences. Training consisted of a series of discrimination and identification tasks for 22 consonants and three vowels in CV context, with feedback, and different materials were used throughout. Overall improvement on

syllables was 11 percentage points, on words was seven percentage points, and on sentences 18 percentage points (over all participants and test conditions), and learning was apparently retained over the 7.5 week interim.

In sum, of these consonant identification speech perception training programs that included a pretest and posttest for sentence recognition, Walden et al. (1981) showed improvements but administered the same pretest and posttest materials, and Lesner et al. (1987) showed no improvements in sentence-level speech recognition when novel materials were used in the posttest. Massaro et al. (1993), though they used a different approach to training and testing than was used in the present study, showed gains slightly greater than those obtained in the present study. In this context, the results from the current auditory-visual vowel identification training program, which showed improved auditory-visual sentence recognition for new materials, including sentences that were not trained, are promising.

Results from the present study also can be compared to research on the effects of training with a tactile device on speech recognition¹. Like the training program in the present study, tactile devices have been designed to assist listeners to better understand speech under difficult listening conditions. De Filippo (1984) examined the effects of four to 10 hours of visual consonant identification training with a tactile device, on lipreading alone versus lipreading with the tactile device. In similarity to the present study, five adults with hearing loss showed gains in sentence recognition of 0 – 15 percentage points when using the tactile device but the aid had no effect on open-set

¹ Kishon-Rabin, Boothroyd, and Hanin (1996) provide an excellent summary of over a dozen published studies of tactile speechreading enhancement for tests of sentence recognition. Practice time in those studies ranged from no practice to 1.5 years of practice; the weighted mean enhancement over studies with deaf adults was 8.7 percentage points, and for normal hearing adults was 6.9 percentage points.

word recognition. Eberhardt et al. (1990) investigated whether vibrotactile presentation of fundamental frequency information acted as an effective supplement to visual speech recognition for normal-hearing adults. Three participants received training under visual-only conditions and 12 received training under conditions of vibrotactile stimulation in addition to lipreading, for approximately 17 hours. Results showed that the vibrotactile stimulation significantly improved speech recognition over vision alone, but the magnitude of improvements was small (an improvement of 4.2 percentage points in words correctly identified) and learning proceeded slowly. Bernstein et al. (1991) investigated sentence recognition following training with or without a vibrotactile vocoder. After 65-70 hours of training it was seen that performance on the Central Institute for the Deaf Everyday Sentences improved significantly from pretest to posttest over all participants, by 9.7 percentage points, but there was no difference in improvement between groups.

The studies by De Filippo (1984), Eberhardt et al. (1990) and Bernstein et al. (1991) were similar to the present study in that phoneme identification training was done (with or without a tactile device), and pretest and posttest materials included the Central Institute for the Deaf Everyday Sentences. Of these, Eberhardt showed little change with training, and De Filippo and Bernstein et al. produced learning of a magnitude similar to that reported in the present study but for training programs longer in duration. While a variety of training programs have produced improvements in the ability to use visual cues in speech recognition, the training program outlined in the present study may have practical utility for adults with hearing loss because many households now have personal computers whereas very few adults have ready access to a vibrotactile device.

V. SUMMARY AND CONCLUSIONS

This study has shown that the vowel identification abilities of normal-hearing adults tested in auditory-visual conditions with noise can be improved with a relatively limited amount of training. All participants in the *vowel trained* group improved vowel identification in monosyllables. Though the word recognition results showed essentially no change from pretest to posttest, sentence recognition demonstrated modest and significant gains for the *vowel trained* participants compared to the *untrained* participants.

Taken together, these results suggest that learning to recognize the fine distinctions between vowels improved auditory-visual speech perception, and that auditory-visual speech recognition for vowels generalized to sentence-level auditory-visual speech perception abilities. Though the increase in sentence-level auditory-visual speech perception ability was not as great as hoped, improvement was in the range of that obtained for other training methods involving visual speech perception (approximately 10 percentage points from pretest to posttest).

Results from this study indicate that auditory-visual speech perception is indeed an ability which may be improved with an appropriate training program. Comparison of results from the current study and results obtained in previous training studies shows that vowels may indeed be useful auditory-visual speech perception training materials. This relatively short-term vowel-based training program yielded promising results, and it is possible that a more intense and/or longer-duration version of the vowel training program described in this experiment, a vowel training program graduated in complexity, or vowel

training with different types of feedback information, may lead to even greater improvements in auditory-visual speech perception ability. As such, the training program described in this experiment may be an effective part of an overall successful approach to auditory-visual speech perception training, and improved speech understanding for adults with hearing loss. Future work will compare results from the current study with a similarly-designed consonant-based auditory-visual speech perception training program, and consonant plus vowel combined training program, in order to provide a better basis for comparison among training methods.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Patrick Zurek and Lynne Bernstein for permission to use stimulus materials, and Drs. Kenneth de Jong, Larry Humes, Jennifer Lentz, and Charles Watson for their many valuable suggestions. This research was supported by NIHDCD-02229 to Indiana University.

REFERENCES

- Ades, A.E. (1977). Theoretical notes: Vowels, consonants, speech, and nonspeech. *Psychological Review*, 84, 524-530.
- Andersson, U., & Lidestam, B. (2005). Bottom-up driving speechreading in a speechreading expert: The case of AA (JK023). *Ear & Hearing*, 26, 214-224.
- Berger, K. (1970). Vowel confusions in auditory-visual speech perception. *The Ohio Journal of Speech and Hearing*, 5, 123-128.
- Bernstein, L.E., Demorest, M.E., Coulter, D.C., & O'Connell, M.P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America*, 90, 2971-2984.
- Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (2000). Speech perception without hearing. *Perception & Psychophysics*, 62, 233-252.
- Bernstein, L., & Eberhardt, S. (1986). *Johns Hopkins lipreading corpus I-II* [Videodisk]. Baltimore, MD: The Johns Hopkins University.
- Binnie, C. (1977). Attitude changes following auditory-visual speech perception training. *Scandinavian Audiology*, 6, 13-19.
- Black, J., O'Reilly, P., & Peck, L. (1963). Self-administered training in lipreading. *Journal of Speech and Hearing Disorders*, 28, 183-186.
- Burkle, Z., Kewley-Port, D., Humes, L., & Lee, J. (2004). Contribution of consonant versus vowel information to sentence intelligibility by normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 115, 2601.
- Cole, R., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. *Proceedings of*

- the 1996 International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, 853-856.
- Davis, H., & Silverman, S. (1970). *Hearing and Deafness (3rd ed.)*. NY: Holt, Rinehart, and Winston.
- De Filippo, C.L. (1984). Laboratory projects in tactile aids to lipreading. *Ear and Hearing*, 5, 211-227.
- Diehl, R.L., Kluender, K.R., Foss, D. J., Parker, E. M., & Gernsbacher, M.A. (1987). Vowels as islands of reliability. *Journal of Memory and Language*, 26, 564-573.
- Eberhardt, S.P., Bernstein, L.E., Demorest, M.E., & Goldstein, M.H. (1990). Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency. *Journal of the Acoustical Society of America*, 88, 1274-1285.
- Erber, N. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12, 423-425.
- Fisher, C. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11, 796-800.
- Gagné, J. (1994). Visual and audiovisual speech perception training: Basic and applied research needs. In J. Gagné & N. Tye-Murray (Eds.), *Research in Audiological Rehabilitation: Current Trends and Future Directions. Journal of the Academy of Rehabilitative Audiology, Monograph Supplement, XXVII*, 133-159.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). *DARPA TIMIT acoustic-phonetic continuous speech corpus* (CD-Rom). National Institute of Standards and Technology.

- Gesi, A., Massaro, D., & Cohen, M. (1992). Discovery and expository method in teaching visual consonant and word identification. *Journal of Speech and Hearing Research, 35*, 1180 – 1188.
- Grant, K., & Walden, B. (1995). Predicting auditory-visual speech recognition in hearing-impaired listeners. Presented at *The XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, 3, 122-129..
- Grant, K., Walden, B., & Seitz, P. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*, 2677-2690.
- Hack, Z., & Erber, N. (1982). Auditory, visual, and auditory-visual perception of vowels by hearing-impaired children. *Journal of Speech and Hearing Research, 25*, 100-107.
- Heider, F., & Heider, G. (1940). An experimental investigation of lipreading. *Psychological Monographs, 52*, 124-153.
- Humes, L., Dirks, D., Bell, T., & Kincaid, G. (1987). Recognition of nonsense syllables by hearing-impaired listeners and by noise-masked normal hearers. *Journal of the Acoustical Society of America, 81*, 765-773.
- Hutton, C. (1959). Combining auditory and visual stimuli in aural rehabilitation. *The Volta Review, 61*, 316 – 319.
- Jackson, P., Montgomery, A., & Binnie, C. (1976). Perceptual dimensions underlying vowel lipreading performance. *The Journal of Speech and Hearing Research, 19*, 796 – 812.

- Kewley-Port, D., Burkle, T., & Lee, J., (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *122*, 2365-2375.
- Kishon-Rabin, L., Boothroyd, A., & Hanin, L. (1996). Speechreading enhancement: A comparison of spatial-tactile display of voice fundamental frequency (F_0) with auditory F_0 . *Journal of the Acoustical Society of America*, *100*, 593-602.
- Lee, H., Rayner, K., & Pollatsek, A. (2001). The relative contribution of consonants and vowels to word recognition during reading. *Journal of Memory and Language*, *44*, 189-205.
- Lesner, S., Sandridge, S., & Kricos, P. (1987). Training influences on visual consonant and sentence recognition. *Ear and Hearing*, *8*, 283-287.
- Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the Speech Code. *Psychological Review*, *74*, 431 – 461.
- MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*, 131-141.
- Mansfield, J.S., Legge, G.E., Luebker, A., & Cunningham, K. (1994). *The MNRead Acuity Charts*. Lighthouse Low Vision Products Inc. New York.
- Massaro, D., Cohen, M., & Gesi, A. (1993). Long-term training, transfer and retention in learning to lipread. *Perception and Psychophysics*, *53*, 549-562.
- McGrath, M., Summerfield, Q., & Brooke, M. (1984). Roles of lips and teeth in lipreading vowels. *Proceedings of the Institute of Acoustics*, *6*, 401-408.
- Mohammed, T., Campbell, R., MacSweeney, M., Milne, E., Hansen, P., & Coleman,

- M. (2005). Speechreading skill and visual movement sensitivity are related in deaf speechreaders. *Perception, 34*, 205-216.
- Montgomery, A., Walden, B., Schwartz, D., & Prosek, R. (1984). Training auditory-visual speech reception in adults with sensorineural hearing loss. *Ear and Hearing, 5*, 30-36.
- O'Neill, J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *The Journal of Speech and Hearing Disorders, 19*, 429-439.
- Prall, J. (1957). Lipreading and hearing aids combine for better comprehension. *The Volta Review, 59*, 64 - 65.
- Richie, C., & Kewley-Port, D. (2003). The effects of auditory and visual vowel training on auditory-visual speech perception performance. *Journal of the Acoustical Society of America, 114*, 2337.
- Richie, C., & Kewley-Port, D. (2005). The effects of auditory-visual vowel and consonant training on auditory-visual speech perception performance. *Journal of the Acoustical Society of America, 117*, 2570.
- Richie, C., Kewley-Port, D., & Coughlin, M. (2003). Discrimination and identification of vowels by young, hearing-impaired adults. *Journal of the Acoustical Society of America, 114*, 2923 – 2933.
- Robert-Ribes, J., Schwartz, J., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America, 103*, 3677 – 3689.
- Sanders, D. (1993). *Management of hearing handicap: Infants to elderly* (3rd ed.). New

- Jersey: Prentice-Hall, Inc.
- Sensimetrics Corporation (2001). *Seeing and hearing speech: Lessons in lipreading and listening* [CD-Rom]. Somerville, MA.
- Squires, S.C., & Dancer, J.E. (1986). Auditory versus visual practice effects in the intelligibility of words in everyday sentences. *The Journal of Auditory Research*, 26, 5-10.
- Studebaker, G.A. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research*, 28, 455-462.
- Sumby, W., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- Tillberg, I., Rönnerberg, J., Svärd, I., & Ahlner, B. (1996). Audio-visual speechreading in a group of hearing aid users; The effects of onset age, handicap age, and degree of hearing loss. *Scandinavian Audiology*, 25, 267-272.
- Tye-Murray, N. (2004). *Foundations of aural rehabilitation: Adults, children and their families* (2nd ed.). San Diego, CA: Dellmar Publishing Co.
- Van Tasell, D., & Hawkins, D. (1981). Effects of guessing strategy on auditory-visual speech perception test scores. *American Annals of the Deaf*, 126, 840-844.
- Walden, B., Erdman, S., Montgomery, A., Schwartz, D., & Prosek, R. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research*, 24, 207-216.
- Walden, B., Prosek, R., Montgomery, A., Scherr, C., & Jones, C. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20, 130-145.

Warren, Y., Dancer, J., Monfils, B., & Pittenger, J. (1989). The practice effect in speechreading distributed over five days: Same versus different CID sentence lists. *The Volta Review*, 91, 321-325.

White, S., Dancer, J., & Burl, N. (1996). Auditory-visual speech perception and auditory-visual speech perception tests: A survey of rehabilitative audiologists. *American Annals of the Deaf*, 141, 236 – 239.

Table 1. Basic protocol for the *untrained* and *vowel trained* participants tested in this study.

Day 1	Day 2-7	Day 8
Both groups	<i>vowel trained</i> group	Both groups
Pretests (no feedback): Vowel identification Word recognition Sentence recognition	Training (with feedback): Vowel identification	Posttests (no feedback): Vowel identification Word recognition Sentence recognition

Table 2. Pretest and posttest vowel identification, for the *untrained* participants and *vowel trained* participants. Performance is expressed as percentage correctly identified vowels, averaged over ten vowels tested.

Participant	Pretest vowel identification	Posttest vowel identification
Untrained		
F1	38	52
F2	43	52
F3	45	45
M1	57	38
M2	33	40
M3	28	40
M4	47	48
Mean	41.6	45
Vowel Trained		
F4	43	67
F5	52	82
F6	22	55
F7	45	63
M5	35	60
M6	35	70
M7	37	70
Mean	38.4	66.7

Table 3. Pretest and posttest word recognition performance, for the *untrained* participants and *vowel trained* participants. Performance is expressed in terms of percent correctly identified words.

Participant	Pretest word recognition for three talkers	Posttest word recognition for three familiar talkers	Posttest word recognition for three unfamiliar talkers
Untrained			
F1	15	10	40
F2	15	10	10
F3	25	30	20
M1	25	30	30
M2	35	30	40
M3	30	20	0
M4	60	50	40
Mean	29.3	25.7	25.7
Vowel Trained			
F4	35	20	20
F5	40	50	40
F6	40	60	40
F7	45	40	50
M5	25	0	40
M6	35	40	40
M7	25	50	30
Mean	35	37.1	37.1

Table 4. Pretest and posttest sentence recognition performance, for the *untrained* participants and *vowel trained* participants. Performance is expressed in terms of keywords correctly identified, in percentage correct.

Participant	Pretest sentence recognition	Posttest sentence recognition
Untrained		
F1	19	14
F2	18	9
F3	22	14
M1	33	36
M2	10	7
M3	6	9
M4	20	21
Mean	18.3	15.7
Vowel Trained		
F4	12	23
F5	33	24
F6	25	49
F7	24	34
M5	11	25
M6	16	33
M7	23	24
Mean	20.6	30.3

FIGURE CAPTIONS

Figure 1. Vowel identification training data for the 7 participants in the *vowel trained* program. Different lines represent the individual participants' performance averaged over the 10 vowels, for each of the 36 training blocks.

Figure 2. Confusion matrices for percent correct identification of the ten vowels in this study, averaged over all participants within a group, for the pretest and posttest.

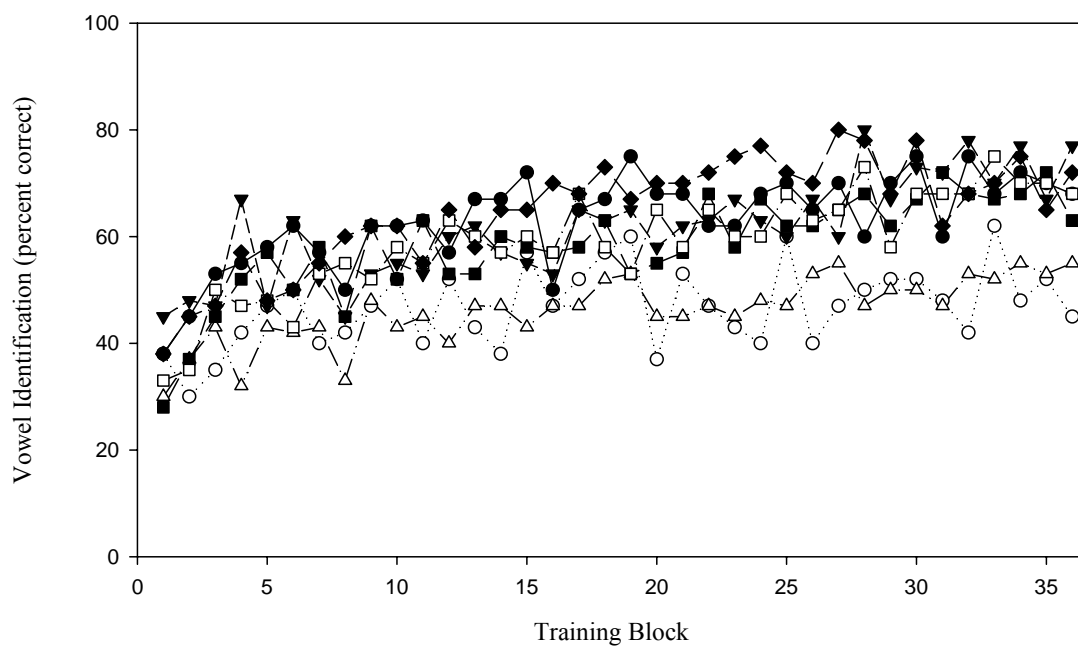


Figure 1

Pretest performance for the *untrained* group

		Stimulus									
		i	ɪ	e	ɛ	æ	ɑ	ʌ	o	ʊ	u
Response	i	37	18	3	2	4	2	4		3	
	ɪ	23	20	9	6	6	5	8		3	
	e	20	18	43	6	20	2				
	ɛ	11	13	11	31	16	14	12			
	æ	3	11	29	27	45	14	22			
	ɑ	3	2	6	10	6	48	20	2		
	ʌ		14		16	2	10	22	7	11	4
	o		2					4	86	6	18
	ʊ	3	2	3			5	6	2	57	25
	u								2	20	57

Pretest performance for the *vowel trained* group

		Stimulus									
		i	ɪ	e	ɛ	æ	ɑ	ʌ	o	ʊ	u
Response	i	43	18	9		2	2	4			
	ɪ	31	21	6	8	4		8		6	
	e	6	9	31	20	20	10	2	2		
	ɛ	11	25	20	37	27	2	2		3	
	æ		2	23	16	41	12	12		3	
	ɑ		2	11	12		50	22	10	14	4
	ʌ	9	14		4	2	14	29	7	6	
	o						5	4	57	3	11
	ʊ		7		2	2	5	12	14	31	32
	u		2			2		4	10	34	54

Posttest performance for the *untrained* group

		Stimulus									
		i	ɪ	e	ɛ	æ	ɑ	ʌ	o	ʊ	u
Response	i	49	16	7		6	2				
	ɪ	26	22	7	2	8	5	10			3
	e	9	8	52	17	18	2				
	ɛ	11	27	12	21	18	5	14			
	æ		16	14	33	37	7	14			
	ɑ			2	17	10	62	29	5		3
	ʌ	6	12	2	7	2	10	27	2	6	
	o				2		2	2	79	23	3
	ʊ				2		5	2	2	46	31
	u							2	12	26	57

Posttest performance for the *vowel trained* group

		Stimulus									
		i	ɪ	e	ɛ	æ	ɑ	ʌ	o	ʊ	u
Response	i	83	14	2		2					
	ɪ	14	65	7	10	6	2	4			
	e	3	6	48	17	18					
	ɛ		8	26	50	4					
	æ		6	17	17	51	2	2			
	ɑ				7	8	64	31			
	ʌ					10	17	61		6	
	o						5		98		3
	ʊ						10			80	20
	u							2	2	14	77

Figure 2